Support Vector Machines

Nipun Batra

July 26, 2025

IIT Gandhinagar

Non-Linearly Separable Data

pik1.png

Data not separate in $\mathbb{R}$

Can we still use SVM?

Yes!

How? Project data to a higher dimensional space.

Non-Lineary Separable Data

pik2.png

Non-Linear Separable Data

pik3.png

Another Example Transformation

pik4.png

Projection/Transformation Function

$\phi : \mathbb{R}^d \to \mathbb{R}^D$

where, $d =$ original dimension

$D =$ new dimension

In our example:

$d = 1; D = 2$

Linear SVM:

Maximize

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \overline{x_i}.\overline{x_j}$$

such that constriants are satisfied.

$$\downarrow$$

Transformation (ϕ)

$$\downarrow$$

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \phi(\overline{x_i}).\phi(\overline{x_j})$$

Trivial Example (Again)

pik5.png

Steps

1. Compute $\phi(x)$ for each point

$$\phi : \mathbb{R}^d \to \mathbb{R}^D$$

Q. If $D >> d$

Both steps are expensive!

Steps

1. Compute $\phi(x)$ for each point

$$\phi : \mathbb{R}^d \rightarrow \mathbb{R}^D$$

2. Compute dot products over $\mathbb{R}^D$ space

Q. If $D >> d$

Both steps are expensive!

Can we compute $K(\bar{x}_i, \bar{x}_j)$

s.t.

$K(\bar{x}_i, \bar{x}_j) = \phi(\bar{x}_i).\phi(\bar{x}_j)$

where,

$K(\bar{x}_i, \bar{x}_j)$ is some function of dot product in original dimension

$\phi(\bar{x}_i).\phi(\bar{x}_j)$ is dot product in high dimensions (after transformation)

Kernel Trick

pik6.png

$\phi(x) = < \sqrt{2}x, x^2 >$
$K(x_i, x_j) = (1 + x_i x_j)^2 - 1$ where $x_i x_j$ is dot product in lower dimensions

$$
\begin{aligned}
(1 + x_i x_j)^2 - 1 &= 1 + 2x_i x_j + x_i^2 x_j^2 - 1 \\
&= < \sqrt{2}x_i, x_i^2 > . < \sqrt{2}x_j, x_j^2 > \\
&= \phi(x_i).\phi(x_j)
\end{aligned}
$$

Kernel Trick

pik7.png

$\phi(x_1) =< -2\sqrt{2}, 4 >; \phi(x_2) =< -\sqrt{2}, 1 >$ Transformation

$\phi(x_1)\phi(x_2) = -2\sqrt{2} \times -\sqrt{2} + 4 \times 1 = 8$ Dot product in 2D

$K(x_1, x_2) = \{1 + (-2) \times (-1)\}^2 - 1$ Dot product in 1D

Kernel Trick

Q) Why did we use dual form?

Kernels again!!

Primal form doesn't allow for the kernel trick

$K(\bar{x}_1, \bar{x}_2)$ in dual and compute $\phi(x)$ and then dot product in D dimensions

Gram Matrix: (Positive Semi-Definite)

$$K(x_i, x_j) = (1 + x_i x_j)^2$$

$$\begin{array}{c|ccccccc}
 & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\
x_1 & 24 & 8 & 0 & 0 & 8 & 24 & 48 \\
x_2 & 8 & 1 & 0 & -1 & 0 & \ldots & \\
x_3 & 0 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
x_4 & 0 & & & & & & \\
x_5 & 8 & & & & & & \\
x_6 & 24 & & & & & & \\
x_7 & 48 & & & & & &
\end{array}$$

3.png

Q) What is $\phi(x)$?

$$K(\bar{x}, \bar{x'}) = \phi(\bar{x})\phi(\bar{x'})$$

$$K(\bar{x}, \bar{x'}) = \left\{ \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} x_1' \\ x_2' \end{bmatrix} \right\}^2 = (x_1x_1' + x_2x_2')^2$$

$$\implies \phi(x) =< x_1^2, \sqrt{2}x_1x_2, x_2^2 >= x_1^2x_1'^2 + x_2^2x_2'^2 + 2x_1x_1'x_2x_2'$$

Another Example

pik9.png

Some Kernels

1. Linear: $\mathsf{K}(\bar{x}_1, \bar{x}_2) = \bar{x}_1 \bar{x}_2$

Some Kernels

1. Linear: $\mathrm{K}(\bar{x}_1, \bar{x}_2) = \bar{x}_1 \bar{x}_2$
2. Polynomial: $\mathrm{K}(\bar{x}_1, \bar{x}_2) = (p + \bar{x}_1 \bar{x}_2)^q$

Some Kernels

1. Linear: $\mathrm{K}(\bar{x}_1, \bar{x}_2) = \bar{x}_1 \bar{x}_2$
2. Polynomial: $\mathrm{K}(\bar{x}_1, \bar{x}_2) = (p + \bar{x}_1 \bar{x}_2)^q$
3. Gaussian: $\mathrm{K}(\bar{x}_1, \bar{x}_2) = e^{-\gamma ||\bar{x}_1 - \bar{x}_2||^2}$ where $\gamma = \frac{1}{2\sigma^2}$ - Also called Radial Basis Function (RBF)

Q) For $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ what space does kernel $\mathrm{K}(\bar{x}, \bar{x'}) = (1 + \bar{x}\bar{x'})^3$ belong to?

$\bar{x} \in \mathbb{R}^2$

$\phi(\bar{x}) \in \mathbb{R}^?$

$$\begin{aligned} K(x,z) &= (1 + x_1 z_1 + x_2 z_2)^3 \\ &= \dots \\ &=< 1, x_1, x_2, x_1^2, x_2^2, x_1^2 x_2, x_1 x_2^2, x_1^3, x_2^3, x_1 x_2 > \end{aligned}$$

10 dimensional?

Q) For $\bar{x} = x$; what space does RBF kernel lie in?

$$
\begin{aligned}
K(x, z) &= e^{-\gamma||x-z||^2} \\
&= e^{-\gamma(x-z)^2}
\end{aligned}
$$

Now:

$$e^{\alpha} = \sum_{n=0}^{\infty} \frac{\alpha^n}{n!}$$

$\therefore e^{-\gamma(x-z)^2}$ is ∞ dimensional!!

SVM: Parametric or Non-Parametric

Q) Is SVM parametric or non-parametric?

SVM: Parametric or Non-Parametric

Q) Is SVM parametric or non-parametric?

Yes and No

Yes $\rightarrow$ Linear kernel or polynomial kernel (form fixed)

No $\rightarrow$ RBF (form changes with data)

RBF is Non-Parametric

$$\hat{y}(x_{test}) = sign(\bar{w}\bar{x}_{test} + b)$$
$$= sign(\sum_{j=1}^{N_{SV}} \alpha_j y_j \bar{x}_j \bar{x}_{test} + b)$$
$$\hat{y}(X_{test}) = sign(\sum_{j=1}^{N} \alpha_j y_j K(\bar{x}_j, \bar{x}_{test}) + b)$$

$\alpha_j = 0$ where $j \neq$ S.V.

RBF is Non-Parametric

Now $K(\bar{x}_j, \bar{x}_{test})$ for RBF is:

$$e^{-\gamma||\bar{x}_j-\bar{x}_{test}||^2}$$

$\therefore$ Hypothesis is a function of “all” train points

Closer $\bar{x}$ is to $\bar{x}_N$; more is it influencing $\hat{y}(\bar{x})$ - hypothesis function

pik10.png

RBF is Non-Parametric

- Now if we add a point to the dataset

RBF is Non-Parametric

- Now if we add a point to the dataset
- Functional form can adapt (similar to KNN)

RBF is Non-Parametric

- Now if we add a point to the dataset
- Functional form can adapt (similar to KNN)
- $\therefore$ SVM with RBF kernel is non-parametric

Interpretation of RBF

- $\hat{y}(x) = sign(\sum \alpha_i y_i e^{-||x-x_i||^2} + b)$

Interpretation of RBF

- $\hat{y}(x) = sign(\sum \alpha_i y_i e^{-||x-x_i||^2} + b)$
- $-||x - x_i||^2$ corresponds to radial term

Interpretation of RBF

- $\hat{y}(x) = sign(\sum \alpha_i y_i e^{-||x-x_i||^2} + b)$
- $-||x - x_i||^2$ corresponds to radial term
- $\sum \alpha_i y_i$ is the activation component

Interpretation of RBF

- $\hat{y}(x) = sign(\sum \alpha_i y_i e^{-||x-x_i||^2} + b)$
- $-||x - x_i||^2$ corresponds to radial term
- $\sum \alpha_i y_i$ is the activation component
- $e^{-||x-x_i||^2}$ is the basis component

RBF: Effect of γ

γ: How far is the influence of a single training sample

pik11.png